7K

УДК 681.3

Б.В. Крыжановский, В.М. Крыжановский
Центр оптико-нейронных технологий НИИСИ РАН, г. Москва, Россия
Vladimir.Krizhanovsky@gmail.com, iont.niisi@gmail.com

Модифицированная q-нарная модель Поттса с бинаризованными синаптическими коэффициентами*

Практическое применение q-нарных моделей Поттса осложняется высокими требованиями к оперативной памяти (необходимо $32N^2q^2$ бит, где N – число нейронов, q – число состояний нейрона). В работе исследуется модифицированная модель Поттса с бинаризованными синаптическими коэффициентами. Процедура бинаризации позволяет в 32 раза уменьшить размер требуемой оперативной памяти (N^2q^2 бит) и более чем в q раз ускорить алгоритм. Ожидалось, что бинаризация приведет к ухудшению распознающих характеристик. Однако анализ показал неожиданные результаты: процедура бинаризации приводит к увеличению объема нейросетевой памяти в 2 раза. Полученные результаты согласуются с проведенными экспериментами.

Введение

На сегодняшний момент все чаще и чаще встают задачи идентификации (классификации) очень большого массива векторов высоких размерностей. Например, при идентификации атак на большие компьютерные сети приходится работать с векторами длиной $N \sim 4000$ и размерностью признака $q \sim 32$.

Одним из классов нейронных сетей, способных решать подобные задачи, являются *векторные нейронные сети*. Наиболее известная из них – спин-стекольная модель Поттса [1], свойства которой достаточно хорошо исследовались методами статистической физики [2-7], а характеристики памяти исследовались, в основном, методами численного моделирования.

Модель Поттса характеризуется большим объемом нейросетевой памяти ($M \sim Nq(q-1)/4\ln Nq$). Однако практическое применение ВНС осложнено высокими требованиями к оперативной памяти (RAM), поскольку для хранения матрицы связей необходимо $\sim 32N^2q^2$ бит. Так для описанной выше задачи необходимо более 16 Гб. Снизить требования к RAM в 32 раза (до N^2q^2 бит) можно бинаризовав синаптические связи нейронной сети (огрубив их до 1 и 0). Тогда для описанной выше задачи вместо 16 Гб потребуется 500 Мб.

Попытка применить процедуру бинаризации к стандартной q-нарной модели Поттса [1] привела к негативному результату: объем памяти резко уменьшился. Поэтому мы применили процедуру бинаризации к *модифицированной* q-нарной модели Поттса. Модели нейронных сетей, подобные этой, исследовались в ряде работ [8-13]. Это так называемые параметрические нейронные сети. Они ориентированы на реализацию в виде оптического устройства. Для них получены достаточно простые аналитические выражения, описывающие эффективность функционирования, объем памяти и помехоустойчивость.

* Работа поддержана грантом РФФИ (06-01-00109).

Анализ исследуемой модифицированной модели Поттса с бинаризованными синаптическими коэффициентами показал неожиданные результаты: объем *нейросетевой памяти* (НП) в результате бинаризации увеличивается в 2 раза, тогда как ожидалось ухудшение распознающих характеристик. Это объясняется тем, что дисперсия входного сигнала нейрона убывает быстрее, чем его среднее значение. Более того, при тех же параметрах N и q рассматриваемая модель превосходит стандартную модель Поттса [1] по быстродействию более чем в q раз. Это существенно для систем реального времени. Полученные результаты согласуются с проведенными экспериментами. Более того, будет показано, что в области $q > 150$, $q < N < q^2$ емкость памяти в результате бинаризации уменьшается. Однако эта область не имеет практического интереса.

Статья имеет следующую структуру. Во втором разделе дается описание исследуемой модифицированной модели. В третьем разделе получены выражения для емкости памяти. В разделе 4 проводится анализ полученных результатов, их сопоставление с экспериментальными данными.

Постановка задачи

Опишем сначала модифицированную модель Поттса. Это полносвязная сеть из N спин-нейронов, имеющих q различных дискретных состояний ($q \geq 2$). Состоянию с номером k ставится в соответствие орт $\mathbf{e}_k$ в пространстве $R^q (k = 1, 2, ..., q)$. Состояние сети в целом описывается N-мерным q-нарным вектором $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, в котором i-й компоненте $\mathbf{x}_i$ соответствует состояние i-го нейрона в паттерне:

$$\mathbf{x}_i \in \{\mathbf{e}_k\}^q. \tag{1}$$

Отличие этой (модифицированной) модели от стандартной модели Поттса [1] заключается в следующем. В модифицированной модели одна из компонент вектора $\mathbf{x}_i$ равна 1, а все остальные равны 0. В стандартной модели Поттса состояние нейрона задается вектором, одна из компонент которого равна $1 - 1/q$, а все остальные равны $-1/q$.

Синаптическая связь между i-м и j-м нейронами в данной модели задается $q \times q$-матрицей $\mathbf{T}_{ij}$, построенной по правилу Хебба на M эталонных паттернах $X_\mu = (\mathbf{x}_{\mu 1}, \mathbf{x}_{\mu 2}, ..., \mathbf{x}_{\mu N})$, $\mu = \overline{1, M}$:

$$\mathbf{T}_{ij} = \sum_{\mu=1}^{M} \mathbf{x}_{\mu i} \mathbf{x}_{\mu j}^{+} . \tag{2}$$

Причем полагается $\mathbf{T}_{ii} = 0$. А локальное поле на i-м нейроне задается в стандартном виде

$$\mathbf{H}_i = \sum_{j \neq i}^{N} \mathbf{T}_{ij} \mathbf{x}_j . \tag{3}$$

Динамика сети определяется естественным образом: i-й спин-нейрон под воздействием локального поля $\mathbf{H}_i$ принимает положение, наиболее близкое к направлению этого поля (состояние нейрона дискретно, поэтому он не может

ориентироваться строго вдоль вектора $\mathbf{H}_i$). Иными словами, нейрон ориентируется вдоль того орта, проекция которого на вектор $\mathbf{H}_i$ максимальна. Соответствующий этому правилу алгоритм заключается в следующем: в каждый момент времени t вычисляются проекции вектора локального поля $\mathbf{H}_i$ на все орты q-мерного пространства, и находится максимальная из этих проекций (пусть, например, это будет проекция на орт $\mathbf{e}_r$). Тогда, состояние i-го спина-нейрона в последующий момент времени задается решающим правилом:

$$\mathbf{x}_i(\mathrm{t}+1) = \mathbf{e}_r . \tag{4}$$

Эта процедура последовательно применяется ко всем нейронам (асинхронная динамика), пока система не конвергирует в стабильное состояние.

Теперь опишем бинаризированную модель. Ее матрица синаптических связей T_{ij} получается процедурой бинаризации исходной Хэббовской матрицы $\mathbf{T}_{ij}$, то есть матричные элементы задаются выражением:

$$\mathsf{T}_{ij}^{\alpha\beta} = \operatorname{sgn} T_{ij}^{\alpha\beta} , \tag{5}$$

где $\alpha, \beta = \overline{1, q}$, $\quad i, j = \overline{1, N}$, а локальное поле принимает вид:

$$\mathbf{h}_i = \sum_{j \neq i}^{N} \mathsf{T}_{ij} \mathbf{x}_j . \tag{6}$$

Динамика бинаризированной модели такая же, как и в исходной модели. Ниже проведено исследование и сравнение свойств этих двух моделей – бинаризированной и исходной (небинаризированной).

Эффективность распознавания эталонов

Оценим объем *нейросетевой памяти* бинаризированной модели. Для этого найдем вероятность того, что записанный эталон является неподвижной точкой. Пусть начальное состояние сети соответствует эталону X_1. Паттерн X_1 будет неподвижной точкой, если для любого из нейронов выполняется условие: проекция локального поля на орт, соответствующий состоянию нейрона в эталонном паттерне, максимальна. Рассмотрим 1-й нейрон и условимся, для простоты, что 1-я компонента X_1 равна $\mathbf{e}_1$. Тогда условие правильной ориентации 1-го нейрона заключается в одновременном выполнении $(q-1)$ неравенств

$$\eta_k = h_1^1 - h_1^k > 0 , \; k = \overline{2, q} , \tag{7}$$

вытекающих из решающего правила (6) и соответствующих тому, что проекция вектора h_1^1 локального поля на орт $\mathbf{e}_1$ больше проекции на любой из остальных $(q-1)$ ортов. Из (5) получим проекции на 1-й и произвольный k-й орт ($k \neq 1$):

$$h_1^1 = N , \tag{8}$$

$$h_1^k = \sum_{i}^{N} \operatorname{sgn}\left(\sum_{\mu \neq 1}^{M} x_{\mu 1}^{k} x_{\mu i}^{\beta_i} \right) , \tag{9}$$

где β_j – номер отличной от 0 компоненты входного вектора $\mathbf{x}_j^{(1)}$.

Вероятность одновременного выполнения $(q-1)$ событий (7)

$$P = \Pr\left[\bigcap_k^{q-1} \eta_k > 0\right] \tag{10}$$

можно вычислить, полагая, что величины η_k случайные гауссовские переменные (не независимые) со средним $<\eta_k>$, дисперсией σ^2 и ковариацией $\mathrm{cov}(\eta_k, \eta_r)$:

$$<\eta_k> = NP_0\,; \tag{11}$$

$$\sigma^2 = NP_0\left[1 + NP_0\left(\tfrac{N-1}{N} P_0^{-\frac{1}{q+1}} - 1\right)\right], \tag{12}$$

$$\mathrm{cov}(\eta_k, \eta_r) = N^2 P_0^2\left(1 - P_0^{-\frac{1}{q^2-1}}\right) \tag{13}$$

Здесь P_0 – это вероятность того, что элемент матрицы связи T_{ij} будет равен нулю

$$P_0 = \left(1 - \frac{1}{q^2}\right)^M. \tag{14}$$

В дальнейшем будем рассматривать случай $q^2 >> 1$ (случай $q \sim 1$ не представляет интереса, поскольку характеристики сети становятся сравнимыми с характеристиками сети Хопфилда).

Наличие параметров (11) – (13) позволяет выписать искомую вероятность в виде стандартного интеграла ошибок для случая мультивариантного гауссовского распределения. Общее выражение мы здесь не приводим ввиду его громоздкости. Опуская промежуточные вычисления, приведем выражение для вероятности P в наиболее интересном случае $P \to 1$. Для этого случая получим справедливое при $\gamma >> 1$ выражение

$$P = 1 - \frac{Nq}{\sqrt{2\pi}\gamma} e^{-\frac{1}{2}\gamma^2}, \tag{15}$$

где $\gamma = <\eta_k> / \sigma$ – так называемое отношение «сигнал/шум». Она является основным показателем надежности распознавания сети: чем больше γ, тем больше вероятность правильного распознавания P и больше объем памяти нейросети.

Объем памяти $\overline{M}$ определим из условия $P \to 1$. В асимптотическом пределе $N \to \infty$ это условие преобразуется к виду:

$$\gamma^2 = 2\ln Nq\,. \tag{16}$$

В случае $N << q$ или $N >> q$ из (17) получим

$$\overline{M} \approx \frac{q^2 N}{2(1 + N/q)\ln Nq}. \tag{17}$$

Получить выражение для емкости памяти при произвольном соотношении величин N и q не представляется возможным в силу трансцендентности уравнения (16). Однако численный расчет показывает, что выражение (17) достаточно хорошо оценивает емкость памяти (с точностью до 20 – 40 %) при любом соотношении величин N и q.

Для сравнения, опуская аналогичные вычисления, приведем оценку емкости памяти исходной небинаризованной сети

$$\overline{M_0} \approx \frac{q^2 N}{4(1+N/q)\ln Nq}. \tag{18}$$

Предваряя дальнейший анализ, сразу отметим, что величина $\overline{M}$ в 2 раза больше величины $\overline{M_0}$. Однако такое соотношение справедливо не на всем диапазоне N и q (рис. 2).

Сравнительный анализ

Проведем сравнительный анализ емкости памяти двух сетей – бинаризированной и исходной моделей. На рис. 1 представлена зависимость величины $\overline{M}$ от размерности задачи N, полученная численным решением уравнения (16). Как видим, емкость памяти бинаризованной сети с ростом размерности сети N быстро нарастает до максимума, а затем медленно логарифмически убывает. Максимальное значение $\overline{M} = M_{\max}$ достигается при $N = N_{\max}$:

$$M_{\max} = \frac{q^{5/2}\ln q}{8\sqrt{2}}, \; N_{\max} = \frac{(2\ln q)^3}{q}\exp\left(\sqrt{q/2}\right). \tag{19}$$

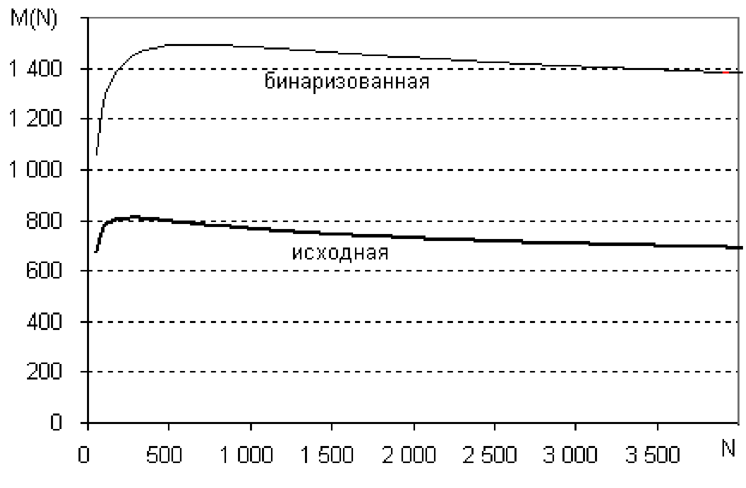


Рисунок 1 – Зависимость емкости памяти от размера сети для бинаризованной (верхняя кривая) и исходной (нижняя кривая) моделей при q = 32

При решении практических задач такое поведение означает следующее: в случае $N > N_{\max}$ можно уменьшить размер эталонов до величины $N = N_{\max}$. При этом повысится емкость памяти, снизятся требования к оперативной памяти и увеличится быстродействие алгоритма.

Для сравнения на рис. 1 приведен график зависимости $\overline{M_0} = \overline{M_0}(N)$, построенный согласно (18). Как видим, величина $\overline{M_0}$ также сначала нарастает, а затем логарифмически спадает. Максимальное значение $\overline{M}_0 = M^0_{\max}$ достигается при $N = N^0_{\max}$:

$$M^0_{\max} = \frac{q^3}{8\ln q}, \quad N^0_{\max} = 2q\ln q \,. \qquad (20)$$

Выражениями (19) и (20) задается максимально достижимая емкость памяти при заданном значении *q*. Кривые на рис. 1 построены при *q* = 32. Как видим, при таком значении *q* память бинаризированной сети приблизительно в 2 раза больше, чем у исходной небинаризированной модели. Такое соотношение выполняется при не слишком больших значениях *q*.

На рис. 2 представлена зависимость отношения $\overline{M} / \overline{M}_0$ от величины *N* при различных значениях *q*. Как видим, практически всюду имеет место соотношение $\overline{M} / \overline{M}_0 > 1$. Более того, $\overline{M} / \overline{M}_0 \sim 2$ при $N < q$ или $N >> q$. Однако при достаточно большой размерности признака ($q > 150$) имеется диапазон изменений *N* ($q < N < q^2$), в котором $\overline{M} / \overline{M}_0 < 1$.

Сравним характеристики сетей, представляющих практический интерес. Это сети, моделирование которых требует не слишком большой оперативной памяти (например, меньше чем 2 Гб). Сети, удовлетворяющие данному условию, принадлежат к области над пунктирной линией рис. 2. Как видим, в этой области процедура бинаризации приводит только к повышению объема нейросетевой памяти.

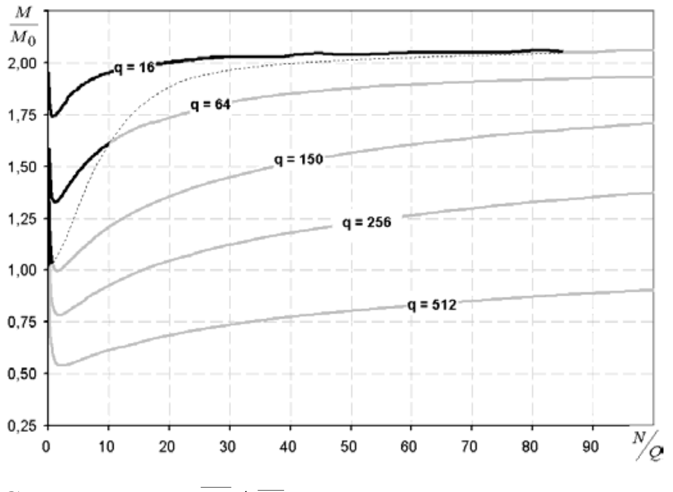


Рисунок 2 – Соотношение $\overline{M} / \overline{M}_0$ при различных значениях параметра *q*. По оси ординат отложена величина *N/q*. Для точек, лежащих под пунктирной линией (кривые, выделенные серым), объем требуемой оперативной памяти, необходимый для моделирования, превышает 2 Гб

Для проверки полученных оценок был проведен ряд экспериментов. Результаты одного из них, проведенного при *N* = 30 и *q* = 10, представлены на рис. 3. Представленные зависимости показывают, что сеть с бинаризованной матрицей связи разрушается при значительно большей загрузке. Кроме того, видим, что емкость памяти бинаризованной модели того же порядка, что и у стандартной

модели Поттса. Однако она превосходит сеть Поттса по быстродействию более чем в q раз (в данном случае в 10 раз) и требует в 32 раза меньше оперативной памяти. Величины параметров сети в различных экспериментах варьировались в пределах $4 \le q \le 128$ и $10 \le N \le 500$. Результаты этих экспериментов показали, что процедура бинаризации всегда приводит к двукратному увеличению объема памяти. Сеть с параметрами ($q > 150$, $N > 32$), при которых бинаризация ухудшает характеристики сети, нам реализовать не удалось в силу ограниченности оперативной памяти компьютера.

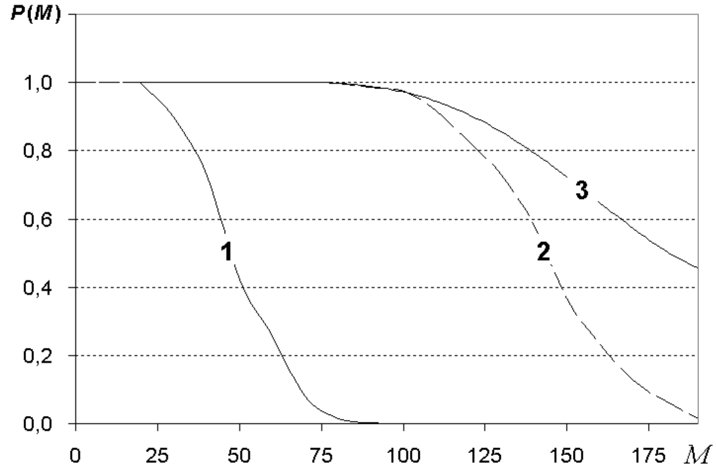


Рисунок 3 – Зависимость доли неподвижных точек P от числа записанных в нейросеть эталонов M для различных моделей: 1 – исходная (небинаризированная) модель; 2 – бинаризированная модель; 3 – модель Поттса. Параметры моделей: $N = 30$ и $q = 10$

Заключение

В работе рассматривалась модифицированная модель Поттса с бинаризированными синаптическими коэффициентами. Показано, что емкость памяти после применения процедуры бинаризации увеличивается вдвое. Это объясняется тем, что дисперсия входного сигнала нейрона убывает быстрее, чем его среднее значение. Для сетей с разумными размерами ($4 \le q \le 128$ и $10 \le N \le 500$) этот результат подтвержден большим числом экспериментов. Более того, рассматриваемая бинаризированная модель сравнима по емкости нейросетевой памяти с стандартной q-нарной моделью Поттса. Однако при этом она превосходит ее по быстродействию более чем в q раз и требует в 32 раза меньше оперативной памяти. Все это делает рассмотренную бинаризированную модель весьма привлекательной для использования в системах идентификации (классификации), работающих в реальном режиме времени с очень большим массивом векторов высоких размерностей.

На основе описанной в работе бинаризованной модели может быть создан q-нарный персептрон для идентификации цветных изображений (эталонов). Этот персептрон позволяет надежно идентифицировать любой из 10^5 эталонов при 75 % искажениях (размер изображения – 52×52 пиксела, число цветовых градаций – 256). Остальные параметры персептрона: число нейронов входного слоя – 2700($= 52 \times 52$), число различных состояний этих нейронов $q = 256$, число нейронов выходного слоя – 2, число состояний выходных нейронов – 325. Для хранения матрицы синаптических

коэффициентов при этом потребуется 56 Мб оперативной памяти, скорость работы алгоритма – 250 мсек (ms). Аналогичная персептронная нейросеть, построенная на поттсовских нейронах, потребует 1,8 Гб оперативной памяти и будет работать в 325 раз медленнее.

Литература

1. Kanter I. Potts-glass models of neural networks. Physical Review A. – 1988. – V. 37(7). – P. 2739-2742.
2. Cook J. The mean-field theory of a Q-state neural network model. Journal of Physics A. – 1989. – № 22. – P. 2000-2012.
3. Vogt H., Zippelius A. Invariant recognition in Potts glass neural networks. Journal of Physics A. – 1992. – № 25. – P. 2209-2226.
4. Bolle D., Dupont P. & Mourik van J. Stability properties of Potts neural networks with biased patterns and low loading. Journal of Physics A. – 1991. – № 24. – P. 1065-1081.
5. Bolle D., Dupont P. & Huyghebaert J. Thermodynamics properties of the q-state Potts-glass neural network. Phys. Rew. A. – 1992. – № 45. – P. 4194-4197.
6. Wu F.Y. The Potts model. Review of Modern Physics. – 1982. – №. 54. – P. 235-268.
7. Nakamura Y., Torii K., Munaka T. Neural-network model composed of multidimensional spin neurons. Phys. Rev. E. –1995. – Vol. 51, № 2. – P. 1538-1546.
8. Kryzhanovsky B.V., Litinskii L.B. and Fonarev A. Parametrical neural network based on the four-wave mixing process. Nuclear Instruments and Methods in Physics Research A. – 2003. – Vol. 502, № 2-3. – P. 517-519.
9. Kryzhanovskii B.V. and Mikaelyan A.L. An associative memory capable of recognizing strongly correlated patterns. Doklady Mathematics. – 2003. – V. 67, № 3. – P. 455-459.
10. Kryzhanovsky B.V., Litinskii L.B., Mikaelian A.L. Vector-neuron models of associative memory. Proc. of Int. Joint Conference on Neural Networks IJCNN-04; Budapest. – 2004. – P. 909-1004.
11. Kryzhanovsky B.V., Mikaelian A.L. and Fonarev A.B. Vector neural net identifying many strongly distorted and correlated patterns. Int. conf on Information Optics and Photonics Technology, Photonics Asia-2004, Beijing-2004. Proc. of SPIE. – 2004. – Vol. 5642. – P. 124-133.
12. Alieva D.I., Kryzhanovsky B.V., Kryzhanovsky V.M., Fonarev A.B. Q-valued neural network as a system of fast identification and pattern recognition. Pattern Recognition and Image Analysis. – 2005. –Vol. 15, № 1. – P. 30-33.
13. Kryzhanovsky B.V., Kryzhanovsky V.M., Mikaelian A.L. and Fonarev A.B. Parametrical Neural Network For Binary Patterns Identification. Optical Memory & Neural Network. – 2005. Vol. 14, № 2. – P. 81-90.

Б.В. Крижановський, В.М. Крижановський

Модифікована q-нарна модель Поттса з бінаризованими синаптичними коефіцієнтами

Практичне застосування q-нарних моделей Поттса ускладнюється високими вимогами до оперативної пам'яті (необхідно $32N^2q^2$ біт, де N – число нейронів, q – число станів нейрона). У роботі досліджується модифікована модель Поттса з бінаризованими синаптичними коефіцієнтами. Процедура бінаризації дозволяє в 32 рази зменшити розмір необхідної оперативної пам'яті (N^2q^2 біт) і більш ніж в q разів прискорити алгоритм. Очікувалося, що бінаризація призведе до погіршення розпізнавальних характеристик. Проте аналіз показав несподівані результати: процедура бінаризації приводить до збільшення об'єму нейромережної пам'яті в 2 рази. Отримані результати узгоджуються з проведеними експериментами.

B.V. Kryzhanovsky, V.M. Kryzhanovsky

Modified q-state Potts Model with Binarized Synaptic Coefficients

Practical applications of q-state Potts models are complicated, as they require very large RAM ($32N^2q^2$ bits, where N is the number of neurons and q is the number of the states of a neuron). In this work we examine a modified Potts model with binarized synaptic coefficients. The procedure of binarization allows one to make the required RAM 32 times smaller (N^2q^2 bits), and the algorithm more than q times faster. One would expect that the binarization worsens the recognizing properties. However, our analysis shows an unexpected result: the binarization procedure leads to the increase of the storage capacity by a factor of 2. The obtained results are in a good agreement with the results of computer simulations.

Статья поступила в редакцию 02.06.2008.